DISCRIMINANT ANALYSIS WITH CATEGORICAL VARIABLES IN MEDICINE

Johannes H a e r t i n g

Institut für Biostatistik und Medizinische Informatik der Martin-
-Luther-Universität Halle-Wittenberg, DDR

Linear discriminant analysis originally developed by R.A. FISHER in the 1930's is a widely applied statistical method in medicine. There are well-equipped computer programs of this method which are implemented in nearly all known statistics program packages.

But a lot of disease entities or disease groups can only be diagnosed with the help of symptoms and signs, that means in statistical terms, with categorical variables. Therefore special approaches are necessary to determine diagnostic allocation rules for categorical data.

We apply several procedures based on the loglinear model which is known from the analysis of multiway contigency tables. But several other procedures are known which work in some instances with comparable results. To begin with, most programs of linear discriminant analysis have the possibility to prescale the variable categories. A second known procedure, which is mostly applied by physicians, is the simple product rule for the marginal probabilities assuming total independence of the variables. At first sight, it's surprising that this procedure works very well in some situations. Later on we will try to explain this circumstance in the view of stability of parameter estimators. A third approach is the so-called logistic discriminant analysis, which estimates the a posteriori probabilities directly with a parametric model.

To evaluate the different approaches correctly an adequate performance criterion is to be chosen. In medicine most diagnostic, therapeutic, and prognostic decisions have to be taken with uncertainty, i.e. in a stochastic environment. Provided that the decision situation is well-defined, the probability of correct classification respectively its complement, the probability of wrong allocations, shortly error rate, is an adequate measure of goodness of the discriminant analysis method. It is intelligible and easy to interpret for medical reasoning. Later on several kinds of error rates and their estimators have to be distinguished.

Let me describe a specified diagnostic situation. An individual /a patient/ is characterized by r categorical variables

$$X = (X_1,\ldots,X_r)' \quad , X_j \text{ with } s_j \text{ categories, } j = 1,\ldots,r$$

which can be either dichotomous /with two possible categories/ or polychotomous. Then the set of all possible states

$$\mathfrak{X} = \{x_1,\ldots,x_s\} \quad \text{has s, the product of all } s_j, \text{ elements.}$$

To ease the following formulae only two populations, disease groups Π_1, Π_2 with given a priori probabilities π_1, π_2 are considered.

The state probabilities in both populations are

$$P(X = x_i \mid \Pi_1) = p_i \ , \quad P(X = x_i \mid \Pi_2) = q_i \quad i = 1,\ldots,s.$$

Then the Bayes-optimal allocation rule, given first by LINHART /1959/, is

$$\begin{aligned} \pi_1 p_i &> \pi_2 q_i && \text{allocation to } \Pi_1 \\ \pi_1 p_i &< \pi_2 q_i && \text{allocation to } \Pi_2 \\ \pi_1 p_i &= \pi_2 q_i && \text{allocation with prob. .5 to } \Pi_1 \text{ resp. } \Pi_2 \end{aligned}$$

This optimal allocation rule yields the smallest possible error rate

$$F^* = F(R^*,f) = \sum_{i=1}^{s} \min \{\pi_1 p_i \ , \ \pi_2 q_i\}$$

In practice, however, the parameters p_i and q_i are unknown. They must be estimated from given samples with known diagnosis or population membership. Maximum likelihood estimates can be achieved as frequencies of individuals in a state in the given sample:

$$\hat{p}_i = \frac{n_{1i}}{n_1} \ , \quad \hat{q}_i = \frac{n_{2i}}{n_2}$$

Substituting them for the unknown parameters in the optimal allocation rule the so-called plug-in-rule is obtained.

When only small or intermediate samples are available this estimated allocation rule often yields disappointing results. This comes from the instability of the estimates becauese of the sparsity of the state counts. Therefore the other mentioned procedures yield sometimes superior results. From our experiences in multiway contingency tables we use in our algorithms loglinear models of different order, especially the model of total independence, the loglinear model with interactions of first order, and the plug-in model from the state counts, also called actuarial model.

The crucial point in the choice of an adequate estimated allocation rule is a good estimator of the rate of wrong allocations with this estimated rule applied for new individuals of unknown disease group membership. This means, the actual error rate

$$F(\hat{R},f) = \sum_{i:\pi_1\hat{p}_i<\pi_2\hat{q}_i} \pi_1 p_i + \sum_{i:\pi_1\hat{p}_i>\pi_2\hat{q}_i} \pi_2 q_i + \sum_{i:\pi_1\hat{p}_i=\pi_2\hat{q}_i} \frac{\pi_1 p_i + \pi_2 q_i}{2}$$

has to be estimated. It depends on the unknown population parameters p_i and q_i.

The simplest way to estimate the actual error rate of a given estimated rule is the resubstitution method. The individuals from the given samples are allocated with the given rule, and the frequency of wrong allocations is counted. The resulting rate is also called apparent error rate:

$$F(\hat{R},\hat{f}) = \sum_{i=1}^{s} \min \{ \pi_1 \hat{p}_i , \pi_2 \hat{q}_i \}$$

But this resubstitution rate has the tendency to give optimistically biased results.

Not only with categorical variables but also with continous variables the following inequality holds (COCHRAN & HOPKINS 1961, HILLS 1966):

$$E\{F(\hat{R},\hat{f})\} \leq F(R^*,f) \leq F(\hat{R},f)$$

It means that the apparent error rate has an optimistic bias as estimator of the optimal error rate and an even greater bias as estimator of the actual error rate.
That's why the apparent error rate is no reliable estimator of the actual error rate. It has to be corrected by an estimator of the bias. A parametric and a nonparametric version are given. To begin with, the bias per state can be expressed as a product of binomial terms. This can be approximated by a normal term. Therefore the total bias can be estimated as sum over normal approximations of the bias per state.

$$E\{F(\hat{R},f) - F(\hat{R},\hat{f})\} \simeq \sum_{i=1}^{s} \hat{\sigma}_i \, \varphi\left(\frac{\pi_1 \hat{p}_i - \pi_2 \hat{q}_i}{\hat{\sigma}_i}\right)$$

$$= \sum_{i=1}^{s} \left(\frac{\pi_1}{n_1} \sqrt{n_{1i}(n_1-n_{1i})} + \frac{\pi_2}{n_2} \sqrt{n_{2i}(n_2-n_{2i})} \right) \times$$

$$\times \varphi\left(\frac{\pi_1 \frac{n_{1i}}{n_1} - \pi_2 \frac{n_{2i}}{n_2}}{\frac{\pi_1}{n_1}\sqrt{n_{1i}(n_1-n_{1i})} + \frac{\pi_2}{n_2}\sqrt{n_{2i}(n_2-n_{2i})}} \right)$$

/$\varphi(.)$: density of the standard normal distribution./

The quite known hold-one-out estimation, for continous variables first proposed by LACHENBRUCH (1968), is a nonparametric version of an estimator of the bais. The estimation will be carried out in a cyclic process. In every cycle one individual will be hold out of the sample, the allocation rule will be estimated from the remaining and will be applied to the single individual. After cyclic repetition for every individual the number of wrong allocations will be counted. This procedure gives a nearly unbiased estimation of the actual error rate. In the case of categorical variables this cyclic process can be brought into a closed formula.

$$\hat{F}_u = F(\hat{R},\hat{f}) + \frac{\pi_1}{n_1}\{ \sum_{A_1} n_{1i} + \frac{1}{2} \sum_{B_1} n_{1i} + \frac{1}{2} \sum_{C_1} n_{1i} \}$$

$$+ \frac{\pi_2}{n_2}\{ \sum_{A_2} n_{2i} + \frac{1}{2} \sum_{B_2} n_{2i} + \frac{1}{2} \sum_{C_2} n_{2i} \}$$

with

$$A_1 = \{\, i: \frac{\pi_2}{\pi_1} \cdot \frac{n_1}{n_2} \cdot n_{2i} < n_{1i} < \frac{\pi_2}{\pi_1} \cdot \frac{n_1-1}{n_2} \cdot n_{2i} + 1 \,\}$$

$$A_2 = \{\, i: \frac{\pi_1}{\pi_2} \cdot \frac{n_2}{n_1} \cdot n_{1i} < n_{2i} < \frac{\pi_1}{\pi_2} \cdot \frac{n_2-1}{n_1} \cdot n_{1i} + 1 \,\}$$

$$B_1 = \{\, i: \frac{n_{2i}}{n_2} < \frac{\pi_1}{\pi_2} \wedge n_{1i} = \frac{\pi_2}{\pi_1} \cdot \frac{n_1-1}{n_2} \cdot n_{2i} + 1 \,\}$$

$$B_2 = \{\, i: \frac{n_{1i}}{n_1} < \frac{\pi_2}{\pi_1} \wedge n_{2i} = \frac{\pi_1}{\pi_2} \cdot \frac{n_2-1}{n_1} \cdot n_{1i} + 1 \,\}$$

$$C_1 = C_2 = \{\, i: \pi_1 \cdot \frac{n_{1i}}{n_1} = \pi_2 \cdot \frac{n_{2i}}{n_2} \,\}$$

Though it looks rather complicated it can be carried out in most cases even with paper and pencil.

The two mentioned methods of estimating the actual error rate are implemented in a combined model choice and variable selection procedure. The principal performance is shown in Figure 1. In order

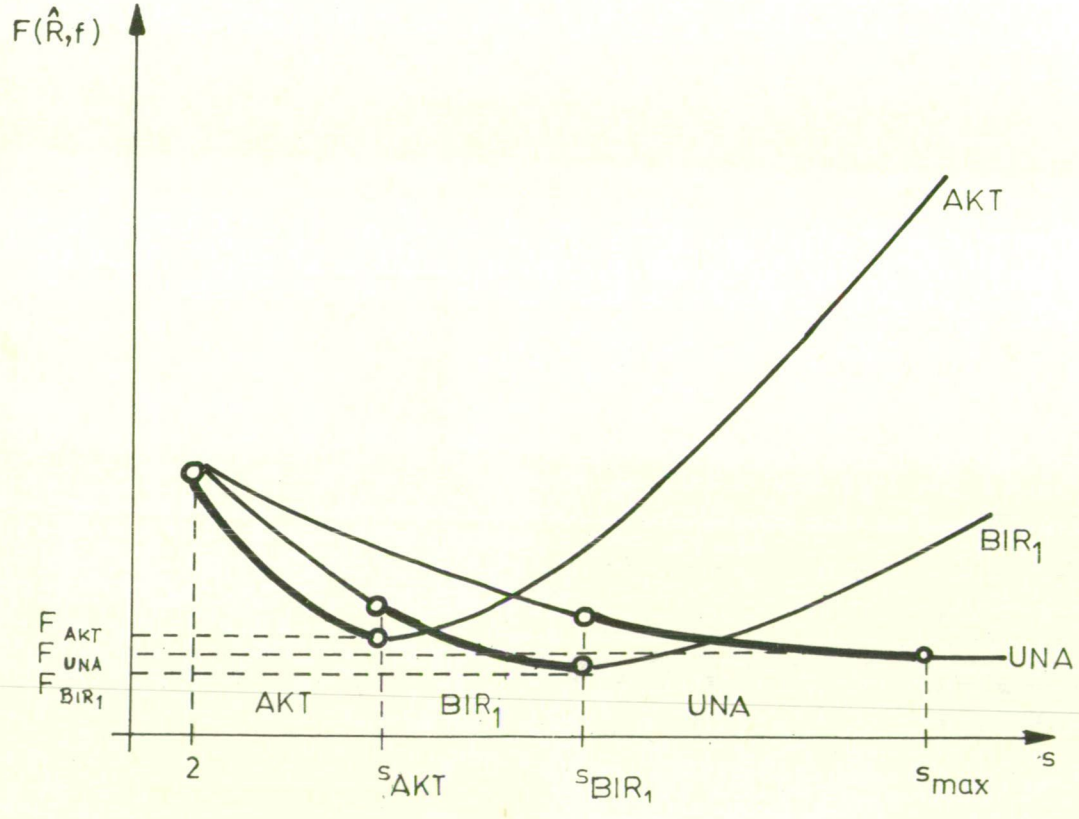

Figure 1.

to limit the computational effort a step by step forward selection and model choice procedure is carried out. We begin with the most complex model /actuarial model/ and the best single variable. Step by step single variables are added up to that point where the actual error rate decreases. Then we change to the loglinear model of first order, add further variables by controlling the actual error rate and eventually change to the independency model and calculate the corresponding estimated actual error rate. We decide for that model and that variable which yields the best estimated actual error rate. The procedure is described in detail in HAERTING (1979). We cannot claim optimality but in our experience it yields rather satisfying results.

References

COCHRAN, W.G. & C.E. HOPKINS: Some classification problems with multivariate qualitative data. Biometrics 17/1961/10-32

HAERTING, J.: Trennverfahren bei qualitativen Merkmalen mit Anwendungen in der algorithmischen medizinischen Diagnostik, Unveröff. Diss. Martin-Luther-Universität Halle-Wittenberg, Halle 1979

HILLS, M.: Allocation rules and their error rates. J. Roy. Statist. Soc. Ser. B 28 /1966/ 1-31

LACHENBRUCH, P.A.: On expected probability of misclassification in discriminant analysis, necessary sample size and a relation with the multiple correlation coefficient. Biometrics 24 /1968/ 323-34

LINHART, H.: Techniques for discriminant analysis with discrete variables. Metrika 2 /1959/ 138-49